# 大数据的热点问题：数据基础设施何去何从？

顾立平[1,2]
1.中国科学院文献情报中心 北京 100190
2.中国科学院经济管理学院信息资源管理系 北京 100190

【摘要】大数据分析和应用涵盖了五个挑战，包括：计算基础设施、数据管理实践、研究人员偏好、各种合作机会和技术掩盖的成本。本文针对这些问题提出了一些切实可行的建议：（1）建立开放的科学平台和数据仓库，促进数据共享和交流；（2） 加强跨学科协作机制，鼓励不同领域的专家参与数据分析和研究；（3） 制定明确的行为准则和规范，以确保数据质量和隐私保护；（4） 利用云计算技术和自动化工具来提高数据处理和分析的效率；（5） 投资大数据领域的教育，培养更多人才，提高整个行业的技术水平。

【关键词】开放科学、开放数据、数据共享、大数据分析、大数据处理、大数据存储
【分类号】G250

# The Hot Issue of Big Data: Where Should Data Infrastructure Go?

Gu Liping[1,2]
1.National Science Library, Chinese Academy of Sciences
2.Department of Information Resource, School of Economic and Management, University of Chinese Academy of Sciences

**[Abstract]** The big data analysis and application, covers five challenges including the computing infrastructure, the data management practices, the researcher preferences, the various collaboration opportunities and the cost obscured by technology. This article proposes some feasible practical suggestions to address these issues as (1) establishing open scientific platforms and data warehouses to facilitate data sharing and communication; (2) strengthening cross-disciplinary collaboration mechanisms and encouraging experts from different fields to participate in data analysis and research; (3) developing clear codes of conduct and specifications to ensure data quality and privacy protection; (4) utilizing cloud computing technologies and automation tools to improve the efficiency of data processing and analysis; (5) investing in education in the field of big data, cultivating more talents, and improving the technical level of the entire industry.
**[Keywords]** Open Science; Open Data; Data Sharing; Big Data Analysis; Big Data Processing; Big Data Store

一、“大数据”落地的挑战

当前大数据分析与应用的热点问题，就是“大数据”怎么落地，对此有五个难点，分别是：计算基础设施、数据管理实践、科研人员的偏好、各种合作的机会以及被技术所掩盖的成本。以下分别论述之。

1.作为科研利器的大数据

大数据是指以往无法处理的大规模、多元、复杂、高维度的数据，其概念于20世纪90年代中期在企业界首次提出，而在学术研究中的应用则始于世纪之交。在过去十年中，大数据及其相关数据科学方法论已成为许多学术领域中重要的研究方法之一。随着数据可用性的增加，大数据已经成为现代科技发展的引擎，广泛应用于大数据研究室、大学图书馆、高性能计算中心、研究项目和研究生项目、个人实验室等。各国科研教育机构已将大数据作为科研利器，不断加速其发展。

2.大数据成为科学研究的关键技术

大数据已从学术研究的边缘逐渐成为越来越多学科研究的核心问题和关键技术。然而，大数据的定义仍然存在争议。尽管数据的规模是重要因素，但大小并非决定性特征。相反，大数据是一个简洁的术语，用于描述以下方面：

（1）利用计算能力和新技术进步的研究项目；

（2）处理、存储和检索数据的用户行为；

（3）整合使用多种工具对研究问题进行探索的任务；

（4）试图结合和解释大规模数据集所实现的技术。。

3.“大数据”落地的外部因素

大数据研究面临着许多外部因素的挑战，这些挑战包括资源需求和保障，人力、资金和政策等方面是许多研究型大学所面临的核心问题。具体而言：

（1）大数据研究需要庞大的计算基础设施，尤其是在存储、共享和分析大规模数据集方面，成本昂贵。

（2）大数据研究中至少需要一个实例记录，并对其进行访问，这需要耗费大量精力。

（3）实验室的大数据研究要求提升大学电网的容量，这可能会占用其他部门的资源。

4.“大数据”落地的内部因素

大数据研究的内部因素也对其实施产生了影响，这些因素包括：

（1）大数据研究需要正式和非正式的合作，涉及到博硕士生、博士后研究人员、科研人员、教授、教职员工、信息技术部门、信息专业人员、图书馆员、法律人员、学术建设委员会以及其他同行的参与。

（2）大数据研究面临着隐藏成本的挑战。虽然大数据基础设施的建设或购买需要相当可观的劳动力投入，但人们常常期望在基础设施完成之后会产生一系列的大数据研究成果，并实现回收成本的效益。然而，实际情况可能并非如此，技术的投入并不能直接带来研究成果的产出。

综上所述，大数据已成为科研利器，已成为许多学科研究的核心问题和关键技术，但面临着资源需求和保障、人力、资金和政策等外部因素的挑战，以及正

式和非正式的合作、隐藏成本等内部因素的影响。

## 二、大数据实践的特点难点

造成上述诸多现象的主要原因，主要有六种互相交错的情况，包括：学科之间的资源争论、复杂数据的管理难度、协作机制的匮乏、分享知识的认识不清、行为规范不够明确、越有效的短期培训造成越不理解根本问题等。

### 1.大数据拉大学科和跨学科之间的张力

大数据研究在学科和跨学科之间存在着张力。尽管大数据研究是一个跨学科的事业，但仍然受到学科组织机构的限制和分散的激励机制的影响。机构的定位、资源的配置、人员的设置和文化结构等因素导致了资金分配的不平衡现象，可能影响不同学科和实践领域参与大数据研究项目的意愿和能力。

一方面，计算机方法的广泛应用，特别是机器学习，推动了数据科学的成熟发展；另一方面，它也导致了科研人员之间的紧张关系，并引发了关于学科观点的争议。

具体而言，统计学、计算机科学、数学和系统工程等学科对于特定领域的数据特征了解不足，难以与具体的学科领域结合。在实践中，这些不同学科的科研人员纷纷建立自己的数据库，称之为大数据，尽管在技术选型、建模思路和科学发现等方面，与真正的大数据还存在较大差距。

### 2.大数据分析的前提是有人管理复杂数据

在当今数据丰富的时代，科研人员通常避免生成新的数据集，而是尽可能利用现有数据。因此，获取可能有用的数据集、清理和组织数据的工作成为了许多大数据项目中最繁重的一部分。

在这种情况下，管理复杂数据的人扮演着至关重要的角色，他们需要具备数据科学、计算机科学、统计学和领域专业知识等多方面的能力。这些人通过使用高级工具和技术，如数据挖掘和机器学习等，帮助科研人员从数据中获取有价值的信息，从而推动大数据分析的发展。

因此，人的角色仍然是大数据分析的前提，尽管技术的发展已经使得处理大规模数据变得更加容易。

### 3.大数据应用依赖协作机制

大数据的应用和分析依赖于广泛的协作机制，涉及到年轻学生、教职员工、同事、客户以及其他机构之外的合作伙伴。实验室在大数据应用和分析的研究中扮演着核心角色，学生（包括本科生和研究生）在实验室中可以对研究过程做出重要贡献。科研人员通常更倾向于使用本地实验室的计算资源，而不是依赖于集中的校园存储和计算选项，包括云计算服务。

然而，实际上，许多学校倾向于投资于信息化建设，并要求科研人员使用这些设施，以证明当初的决策的正确性。如果使用过程不顺利得到认可，进一步的培训或优惠价格等措施将被采取。

这种情况表明，大数据应用的成功离不开各方之间的协调与合作，并需要持续的资源支持和培训。

4.大数据文化的本质是分享知识

尽管同行评审的论文仍然是学术交流中最具激励性的形式，但科研人员应该广泛致力于公开分享研究成果，包括数据和代码。然而，学术分享的实践超出了仅满足 FAIR 原则（可查找性、可访问性、互操作性和可重用性）、开放获取的、正式共享的数据知识库的范畴。

存在一种障碍，即科研人员认为许多数据要么是衍生的、低质量的，要么是从各种来源收集的，不适合公开共享。这种观念限制了大数据文化的本质，即分享知识的精神。

为了促进大数据文化的发展，科研人员应该克服这些障碍，鼓励和支持数据的广泛共享，以促进科学研究的透明度和可重复性。

5.科研人员的行为规范

尽管国家对于科研诚信和学术道德的建设采取了许多政策和公布科研失信案例，但大数据研究的伦理层面仍然存在争议。科研人员在确定科研行为的最佳实践方面仍存在不确定性。

尽管法律法规和管理规则受到重视，但一些科研人员担心这些规定是否能够很好地适应新发展的、不断发展的、基于大数据的研究方法。

因此，对于大数据研究的伦理规范的制定和执行需要更加深入的讨论和研究，以确保科研行为的规范性和合理性，并适应不断变化的科研环境。

6.许多大数据培训的盲点

尽管科研人员倾向于采用非正式的训练方法，如互联网教程或基于实践案例的大数据方法等，这些方法对于解决实际问题具有一定的效果。然而，这种培训方法存在潜在的盲点，尤其是对于大数据领域的基础知识的掌握。基础知识的欠缺可能会在学术研究中导致问题。因此，为了有效地利用大数据技术，科研人员需要接受系统性的大数据培训，以获得必要的基础知识，并在此基础上进一步掌握高级技术和方法。这需要教育机构和社会各界共同努力，提供适合不同层次和需求的大数据培训课程和资源，以培养更多的专业人才。

综上所述，大数据研究在学科和跨学科之间存在着张力，主要受到学科组织机构的限制、分散的激励机制以及资源分配的不平衡等因素的影响。科研人员在处理复杂数据时，管理复杂数据的人扮演着至关重要的角色，他们需要具备多方面的能力。大数据的应用和分析依赖于广泛的协作机制，需要各方之间的协调与合作。科研人员应该广泛致力于公开分享研究成果，以促进科学研究的透明度和可重复性。尽管国家对于科研诚信和学术道德的建设采取了许多政策，但大数据研究的伦理层面仍然存在争议。许多大数据培训的盲点，科研人员需要接受系统性的大数据培训，以获得必要的基础知识，并在此基础上进一步掌握高级技术和方法。

## 三、应对措施

综上所述，本文提出以下七点建议，旨在推进科研院所和高等学校，解决大数据分析与应用方面的落地问题，包括大数据服务供应商在内，如何在短期有效和长期深耕之间，取得合适平衡的策略。如此一来，可以促使大数据如何落地甚

至成为助力科研、助力产业、助力就业的有效手段，更进一步提升我国大数据分析与应用的推广。

**1.科研教育机构应采取以下措施来推动大数据研究的发展：**

（1）定期对校园大数据基础设施进行系统评估，并制定协议，绘制信息技术路线图，以确保数据存储需求与功能的匹配。

（2）组建工作组，包括图书馆、高性能计算、科研领域的资源（如各类大小数据中心）、科技处（或业务处）以及与其他单位协调的支持服务，以促进协同合作和资源共享。

（3）发展正式的数据服务和资源目录，并向科研人员分发，以便他们更好地利用和管理大数据资源。

（4）评估当前科研评价与科研诚信的标准，以确保其能够充分反映大数据研究所涉及的道德和隐私问题。

（5）寻找机会，为资源不足的领域提供支持，包括人文艺术学科、定性的社会科学和一些专业学科，以促进这些领域在大数据研究方面的发展。这可以通过开展专门的培训、提供相关设施和资源等方式实现。

**2.为促进大数据的信息化项目的发展，可采取以下措施：**

（1）增加资助青年科研人员的项目，包括在数据科学和编程方面的实践、推广和专门研究，以培养更多的专业人才，推动大数据研究的发展。

（2）奖励对大数据研究做出贡献的工作，以激励科研人员积极投身于大数据领域的研究和创新。

（3）鼓励资助资金获得者，支持那些几乎没有机会获得外部资助的领域的工作，以促进大数据研究在各个领域的普及和应用。

（4）与其他机构建立联盟关系，建立长期数据存储和计算能力，以提高数据的安全性和可持续性，为大数据研究提供更好的基础设施支持。

（5）为科研人员开发人员和项目管理培训，以提高其技术和管理能力，同时表彰那些本质上是大数据研究协作的工作，促进协作与交流，推动大数据研究的发展。

**3.为推动大数据研究的发展，学院和研究部门可以采取以下措施：**

（1）投资于进一步嵌入数据科学、数据管理、统计和计算流程，为科研人员提供相关专业知识，以协助大数据研究的进行。

（2）在博士项目中，特别是 STEM 领域的博士项目中，应寻求整合机器学习方法、数据科学和编程的机会，并至少将这些内容纳入博士课程中。

（3）研究部门应考虑培养人才的素质教育课程，使研究人员至少了解大数据研究并能为之做出贡献。

（4）修订晋升和任期标准，以确认组织良好的数据和代码共享是一项重要的研究成果，从而鼓励科研人员在数据和代码的管理与共享方面做出努力。

（5）培养团队成员在元数据创建、数据管理和数据管理方面的专业知识，以及数据分析和数据可视化的能力，以提高团队在大数据研究中的综合能力。。

**4.为了促进大数据研究的发展，图书馆可以采取以下措施：**

（1）创建和更新特定科研社区感兴趣的数据集指南，以帮助研究人员快速找到

其需要的数据集资源。
（2）为购买订阅数据集分配额外资源，与其他学术图书馆合作，以降低成本，为科研人员提供更多的数据资源。
（3）增加现有数据研究管理服务的推广活动，这些服务是为科研人员所作，需求量很大。通过加强宣传和推广，可以让更多的科研人员了解到这些服务。
（4）在可行的情况下，扩大一对一咨询服务或提供按需研讨会，根据特定研究群体的需求量身定制，以更好地满足科研人员对数据研究管理的需求。
（5）提高机构知识库的存储能力，并将其推向科研人员。通过提高机构知识库的存储能力，可以更好地存储和管理科研人员的数据资源，提高数据的可访问性和可重复性。

**5.为推动大数据研究的发展，科研资助机构可以采取以下措施：**
（1）评估受资助者当前是否需要法律和道德指导，以满足与大数据研究相关的新兴道德和隐私问题。这将确保受资助者在进行大数据研究时能够充分考虑法律和道德问题，保护个人隐私和数据安全。
（2）制定支持与考核机制，为长期的大数据基础设施的维护成本提供资金。这将确保大数据基础设施的可持续发展，并提供资金支持来确保其正常运行和维护。
（3）继续支持数据知识库的稳健发展。资助机构可以提供资金和技术支持，以帮助建立和维护数据知识库，为科研人员提供可靠的数据资源。
评估现有的代码和数据共享法规在多大程度上为科研人员在处理专有的、机密的、敏感的、低质量的数据时提供正确的指导。这将有助于确保（4）科研人员在共享数据和代码时能够遵守相关法规，同时保护数据的安全和质量。
（5）在科研项目计划评估中，制定有组织的数据和代码共享的评估指标、指南和指导。这将帮助科研人员在项目计划中考虑数据和代码共享的重要性，并提供指导来确保数据和代码的有效共享和可重复性。

**6.为促进学术社区对大数据研究的支持和发展，以下措施可以考虑：**
（1）阐明基于学科的研究伦理观点。学术社区应明确各学科领域对于大数据研究伦理的观点和要求，以确保研究人员在进行大数据研究时遵循适当的伦理原则和规范。
（2）在学术会议和出版物上，鼓励对于开放科学的价值进行详细讨论。学术社区应通过会议和出版物提供平台，促进研究人员对于开放科学的价值进行深入探讨和交流，推动共享数据和代码的实践。
（3）面向科研人员、资助机构、出版团体以及其他利益相关者，阐明数据处理政策和存储标准。学术社区应制定明确的数据处理政策和存储标准，向科研人员、资助机构、出版团体等相关方提供指导，以确保数据的安全、可访问性和可重复性。
（4）鼓励各学术部门在晋升和聘任标准上，制定有组织的数据和代码共享的评估指标。学术部门应考虑将数据和代码共享作为评估科研人员晋升和聘任的标准之一，鼓励科研人员积极参与数据和代码的共享和开放科学实践。
（5）为科研人员举办论坛、研讨会、专题讨论会等活动，提供跨学科共享和数据密集型研究的机会。学术社区应组织各种形式的活动，为科研人员提供交流和

合作的平台，促进跨学科的共享和数据密集型研究的发展。

**7.为提高大数据服务供应商的服务质量和推动大数据研究的发展，以下措施可以考虑：**

（1）增强订阅数据库的元数据。大数据服务供应商应增强订阅数据库的元数据，以提高数据的可搜索性和可发现性，促进数据的共享和重复利用。

（2）与图书馆协调，提供数据集以及读者使用许可。大数据服务供应商应与图书馆协调，制定数据集的供应和使用许可协议，以确保数据的合法使用和保护数据提供商的权益。

（3）提供数据封包及其使用许可证，促使数据集可供科研院所和研究型大学进行访问获取，并且价格合理。大数据服务供应商应提供数据封包和使用许可证，以便科研院所和研究型大学可以访问和获取数据集，并确保数据的价格合理和可负担。

（4）科研社区与科研院所一同协商讨论一个云存储选项的固定价格。大数据服务供应商应与科研社区和科研院所协商讨论一个云存储选项的固定价格，以便科研人员可以更便捷地存储和共享数据。

（5）提供个性化咨询服务，协助科研人员进行特定领域的编码任务和数据管理。大数据服务供应商应提供个性化咨询服务，帮助科研人员解决特定领域的编码任务和数据管理问题，提高数据的质量和可重复性。

## 四、结语

当前大数据分析与应用的热点问题，就是“大数据”怎么落地，对此有五个难点，分别是：计算基础设施、数据管理实践、科研人员的偏好、各种合作的机会以及被技术所掩盖的成本。造成这些现象的主要原因，包括：学科之间的资源争论、复杂数据的管理难度、协作机制的匮乏、分享知识的认识不清、行为规范不够明确、越有效的短期培训造成越不理解根本问题等的六个原因。故此，本文针对科研教育机构、科研资助机构、科学学会、学院以及研究部门、信息化项目、图书馆、大数据服务供应商等，各自提出五项可操作的实践建议，以期共同解决这些热点问题。抛砖引玉，是以为文。

## 参考文献

[1] 顾立平, 张潇月. 分布式大数据资产权益管理问题与对策[J]. 农业图书情报学报, 2023, 1(35):39-55.

[2] Dylan Ruediger 等.Big Data Infrastructure at the Crossroads Support Needs and Challenges for Universities[EB/OL].[2021-12-01]https://doi.org/10.18665/sr.316121